A quick review:

After reading the paper by Keiser et al. (2007) describing the Similarity ensemble approach (SEA) it appeared that in the process of creating the statistical model a few steps were made for which the reasoning behind was lacking or not provided. This drove us to try and recreate a new model.

Methods:

All ligands and drugs fingerprints, target (protein) ligand groups were provided by Keiser from a paper published in 2012 testing SEA predictions in-vitro.

Sampling process:

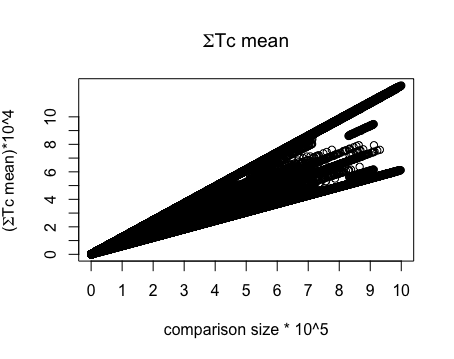
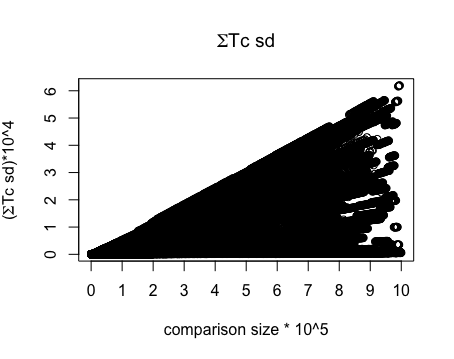
1. Ligand sets of sizes ranging from 1 to 1,000 were randomly selected from a pool of 163,547 ligands.
2. For each two ligand sets, A and B, the Tanimoto coefficient score (Tc) was calculated for each pair of ligands one from set A and the second from set B. The sum of Tc was saved so a tuple was created : (, , ).

The sampling process was repeated 10 times.

Analysis done:

First since SEA sampling was done using “across logarithmic set size intervals in the range of 10 to 1,000 molecules” (it is worth noting that although there were no sets of a single ligand in the sampling process, the statistical model is being used in Keiser et al. (2012) to predict the statistical significance when a single ligand is checked against another set of ligands which sometimes results in a comparison size less than 100). It was interesting to see if the same results would be achieved when using set sizes ranging from 1 to 1,000 (using set size interval of 1).

Using the data sampled a mapping between and was calculated in order to replicated the sampling data produced by Keiser et al. for creating the SEA statistical model.

The results are shown in the following two figures: 

To the left is a plot showing the mean of for each comparison size (which equals in previous notation). When fitting a linear model to mean the model has an R-squared value of 0.8421.

To the right is a plot showing the standard deviation (SD) of for each comparison size.

When fitting a linear model to SD the model has an R-squared value of 0.1325. It is easy to see that trying to fit any model to that data would not produce usable results (In the paper describing SEA a nonlinear fit was computed).

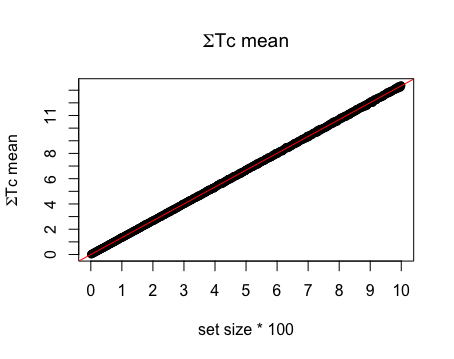
The idea proposed was that the results of the same comparison size differ based on the sizes of the sets used in the comparison. For example, when comparing two sets one of one ligand and the other of 400 ligands the results would differ from the results we would receive when comparing two sets of 20 ligands although the number of comparisons is equal.

We can look at it as if the results of the model are based on the size of smaller of the two sets.

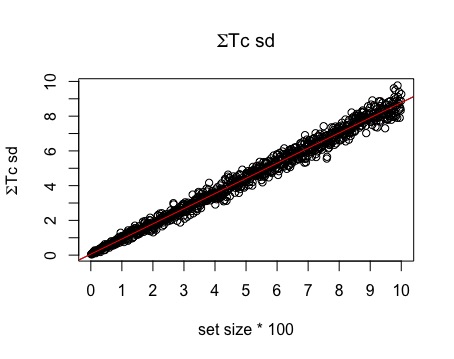
To test this, the data was grouped based on the size of the smaller set of the two used for the comparison.

The results shown below are for a small set of one ligand only.

The following figure shows the relation between the mean of to the size of the second set in the comparison (in the notation used before this plot shows , .



When fitting a linear model to mean the model has an R-squared value of 0.999.

The following figure shows the relation between the SD of to the size of the second set in the comparison (in the notation used before this plot shows , .

When fitting a linear model to SD the model has an R-squared value of 0.9868.

In Keiser et al. (2012) SEA predictions were tested across a number of ligands and proteins. In order to check if we are on the right track I’ve decided to calculate normalized z-scores (ZS) in the same manner they were calculated in Keiser et al. (2007), plot a histogram of the values and mark in vertical lines the ZS for positive SEA predictions.

Each target’s ZS (such as Sertraline) is calculated against two target proteins groups, the first is human only and the second is all organisms for which binding information is available.

The reason I’ve added this plot is I have a feeling we are dealing with two populations, the right most of ZS which belongs to ligands that will bind to tested targets and the left most with of ZS which belongs to ligands that will not bind to tested targets.

